DISAMBIGUATION OF TDOA ESTIMATES IN MULTI-PATH MULTI-SOURCE ENVIRONMENTS (DATEMM)

Jan Scheuing and Bin Yang

Chair of System Theory and Signal Processing
University of Stuttgart, Germany

ABSTRACT

One major problem of time delay estimation for acoustic localization in multi-source reverberant environments is the ambiguity in identifying out of many peaks of generalized cross-correlation the desired time differences of arrival (TDOAs) caused by direct paths and in assigning them correctly to individual sources. In this paper, we propose a novel geometrically motivated approach "Disambiguation of TDOA estimates in multi-path multi-source environments" (DATEMM). It utilizes additional information from the auto-correlation of sensor signals and a zero TDOA sum condition to suppress spurious TDOA estimates. Furthermore, this method can be used as an add-on module to improve the robustness of any existing TDOA estimation method.

1. INTRODUCTION

The position of an acoustic source in a room is usually estimated from signals of a microphone array by applying a multi-stage localization method, mainly consisting of some preprocessing like e.g. VAD, time delay estimation, and estimation of the geometric position. Most approaches up to now work well for one source in a less reverberant environment. The underlying signal model assumes that the source signal $s(t)$ propagates on a direct path with delay τ_i and is received by two sensors $x_i(t) = h_i s(t-\tau_i)$ $(i = 1, 2)$. By generalized cross-correlation [1] of $x_1(t)$ and $x_2(t)$, the TDOA $\tau_1 - \tau_2$ is estimated. Using several sensors with each sensor pair contributing a TDOA measurement, the source position may be estimated e.g. according to [2, 3, 4]. However, for multiple simultaneous speakers or in a reverberant environment the results are not satisfactory.

Recent time delay estimators try to measure the (single-source) room impulse response [5] or utilize a multi-channel cross-correlation [6] in order to cope with the multi-path problem. Only a few articles address the multi-source problem: [7] uses subspace methods, [8] tracks individual sources while they are overlapping, [9] extends a Viterbi search-based system for adaptive beamforming, and [10] is based on blind channel identification.

This paper presents some new ideas for identifying direct-path time-delays and for assigning them to different simultaneous sources in a multi-source and multi-path environment in time domain. We first describe the signal model in section 2. Motivated by an ambiguity analysis of time delay estimation in section 3, we propose a new algorithm for disambiguation of TDOA estimates in section 4. Finally, section 5 presents some simulation results.

2. MULTI-SOURCE MULTI-PATH MODEL

Let us consider a room with N sources and M sensors. A model of a real acoustic environment must take reverberation of the room into account. The sensor signals $x_i(t)$ may be expressed for given source signals $s_a(t)$ in the noise-free case as

$$x_i(t) = \sum_{a=1}^{N} (h_{a,i} * s_a)(t) \quad (i = 1, \dots, M) \tag{1}$$

where $*$ denotes convolution and $h_{a,i}(t)$ is the room impulse response between the a-th source and the i-th sensor. We assume that each room impulse response $h_{a,i}(t)$ consists of a finite number $L_{a,i}$ of significant paths, where the μ-th path is described by its amplitude $h_{a,i,\mu}$ and delay $\tau_{a,i,\mu}$. The delay $\tau_{a,i,\mu}$ is assumed to increase in μ. Therefore, $\mu = 0$ represents the direct path and $\mu \geq 1$ stand for all echo paths. Non-significant paths, noise, and directivity of sensors and sources are neglected in this paper. Hence the signal model becomes

$$x_i(t) = \sum_{a=1}^{N} \sum_{\mu=0}^{L_{a,i}-1} h_{a,i,\mu} s_a(t - \tau_{a,i,\mu}). \tag{2}$$

If we consider one source or one path per source, we drop the corresponding index a or μ in (2), respectively.

In the following, the auto-correlation function of $x_i(t)$ is denoted by

$$r_i(t) = \mathrm{E}\left[x_i(t + t_0)x_i(t_0)\right]. \tag{3}$$

Its local extrema will occur in distance $t_{a,i,\mu\nu} = \tau_{a,i,\mu} - \tau_{a,i,\nu}$ symmetrically to the origin. The cross-correlation function of two sensor signals $x_i(t)$ and $x_j(t)$

$$r_{ij}(t) = \mathrm{E}\left[x_i(t + t_0)x_j(t_0)\right] \tag{4}$$

will show peaks of amplitude $g_{a,ij,\mu\nu}$ at $t_{a,ij,\mu\nu} = \tau_{a,i,\mu} - \tau_{a,j,\nu}$.

1-4244-0469-X/06/$20.00 ©2006 IEEE

3. AMBIGUITY OF TDOA ESTIMATES

Ambiguity of TDOA estimates arises when we obtain multiple maxima in the cross-correlation. In this case, we don't know which TDOA is the correct one for a particular source and a selected sensor pair. There are three obvious reasons for this phenomenon:

- reverberations
- multiple sources
- non-white source signals

Below we analyse the ambiguity for different cases.

3.1. One white source and multiple paths

For a single source, the two sensor signals

$$x_i(t) = \sum_{\mu=0}^{L_i-1} h_{i,\mu} s(t - \tau_{i,\mu}) \quad (i = 1, 2) \tag{5}$$

each contain L_i paths with delay $\tau_{i,\mu}$. If $s(t)$ is white, the cross-correlation of $x_1(t)$ and $x_2(t)$ returns at maximum $L_1 L_2$ local extrema at $t_{12,\mu\nu}$ $(0 \le \mu < L_1,\, 0 \le \nu < L_2)$.

For TDOA based localization, we are only interested in the TDOA of the two direct paths $t_{12,00} = \tau_{1,0} - \tau_{2,0}$. But which peak in $|r_{12}(t)|$ corresponds to $t_{12,00}$? This *multi-path ambiguity* is caused by reverberations.

3.2. Multiple white sources and only direct paths

Regarding N white and uncorrelated sources $s_a(t)$ in an anechoic room, where only the direct paths from source a to sensor i with delay $\tau_{a,i}$ contribute to the sensor signals

$$x_i(t) = \sum_{a=1}^{N} h_{a,i} s_a(t - \tau_{a,i}) \quad (i = 1, 2)\,, \tag{6}$$

the cross-correlation $r_{12}(t)$ will show at maximum N local extrema at $t_{a,12}$ $(1 \le a \le N)$.

But which peak corresponds to which source? Resolving this *multiple-source ambiguity* is important for localization, because we have to assign each TDOA to one source and consider all TDOAs of that particular source together to estimate its geometric position.

3.3. One speech source and only direct paths

Speech signals consist of unvoiced (quasi-random) and voiced (quasi-periodic) parts. A single periodic source in an anechoic room will cause many local extrema at periodic intervals in the cross-correlation $r_{12}(t)$. So which peak corresponds to the desired TDOA? To avoid this *periodic ambiguity*, the sensor signals are usually prewhitened before cross-correlation (e.g. GCC-PHAT [1]). This has about the same effect like cross-correlating white sources.

3.4. Multiple sources and multiple paths

In the following we consider N uncorrelated source signals and M pre-whitened sensor signals according to (2). The cross-correlation $r_{ij}(t)$ between sensor i and j can show at maximum $\sum_{a=1}^{N} L_{a,i} L_{a,j}$ local extrema at $t_{a,ij,\mu\nu}$. What we need for each sensor pair, however, are N TDOAs $t_{a,ij,00}$ caused by the direct paths only. The disambiguation has thus two tasks:

- identify the TDOAs of direct paths $(\mu = \nu = 0)$
- assign direct path TDOAs to individual sources $(a = 1, \ldots, N)$

4. DISAMBIGUATION

Our disambiguation approach DATEMM is based on the following four observations:

A1) Raster match:
By exploiting the peak positions of the auto-correlation function, the extrema positions in the cross-correlation always appear in a certain *raster*: a set of time marks with known distances between them. Fig. 1 shows a simple example with one source, two sensors, and two paths per sensor. The raster in the cross-correlation consists of four time marks whose distances are known from the peak positions of the auto-correlations. By finding this raster in the cross-correlation, its absolute position determines the desired TDOA $t_{ij,00}$ of direct paths.

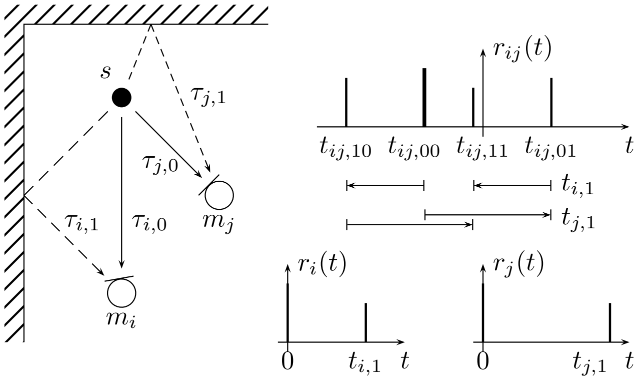

Fig. 1. Raster for one source ($M = 2$, $L_i = 2$, $L_j = 2$)

A2) Zero TDOA sum:
For any subset of $\tilde{M}$ sensors, the sum of TDOAs

$$t_{a,12,\mu_1\mu_2} + \cdots + t_{a,\tilde{M}-1\,\tilde{M},\mu_{\tilde{M}-1}\mu_{\tilde{M}}} + t_{a,\tilde{M}1,\mu_{\tilde{M}}\mu_1}$$

is zero [3], provided that all $\tilde{M}$ TDOAs are estimated for the same source a and the same paths μ_i to sensors $i = 1, \ldots, \tilde{M}$. If this sum deviates from zero, the TDOAs don't stem from the same source and/or paths.

A3) Array size limitations:
If we know the sensor array geometry, we can com-

pute an upper bound for $|t_{a,ij,00}|$ in advance and discard larger TDOA estimates.

A4) Positive extremum for direct path TDOA:
In acoustic localization, the direct path amplitudes $h_{a,i,0}$ are always positive. Hence the amplitude of $r_{ij}(t)$ corresponding to $t_{a,ij,00}$ is positive, as well.

DATEMM is structured to three levels, motivated by the number of sensors involved: pair-level, triple-level and array-level. Besides reduction of ambiguity in TDOA estimates, we also provide quality measures for TDOA estimates at each level. They reflect the reliability of individual TDOA estimates for different sources and will be useful for the estimation of the number of sources N, if it is unknown. Below we sketch the main ideas of DATEMM without implementation details.

4.1. Direct path detection by exploiting auto-correlation

The first step at the pair-level is to detect the direct path TDOA $t_{a,ij,00}$ from the cross-correlation $r_{ij}(t)$ by additionally exploiting auto-correlations $r_i(t)$ and $r_j(t)$. For this purpose, we first extract the relevant peaks from both cross- and auto-correlations, which, hopefully, include the direct paths of all sources. For a practical implementation, we use a given number of most dominant peaks, combined with the condition that the corresponding peaks exceed a certain threshold in magnitude. We use the cross-correlation amplitude $g_{a,ij,\mu\nu}$ as the initial quality of TDOA $t_{a,ij,\mu\nu}$. This value will be increased or decreased during the subsequent steps. We further assume that the direct path is always involved in all selected auto-correlation peaks.[1].

We now consider two TDOAs $t_{a,ij,\mu_1\nu_1}$ and $t_{a,ij,\mu_2\nu_2}$ resulting from the same source a where the paths to sensor i are common ($\mu_1 = \mu_2 = \mu$) and one of the paths to sensor j is a direct path ($\nu_1 = 0$ or $\nu_2 = 0$). Clearly, the distance $|t_{a,ij,\mu\nu_1} - t_{a,ij,\mu\nu_2}|$ can be found as the position of a peak in the auto-correlation of sensor j, see Fig. 1. As the direct path is always the shortest, we can also determine the sign of the above difference and hence identify whether ν_1 or ν_2 is the direct path. Similarly, if $\nu_1 = \nu_2 = \nu$ and one of the two paths to sensor i is a direct path, the TDOA difference $t_{a,ij,\mu_1\nu} - t_{a,ij,\mu_2\nu}$ will match to a peak in $r_i(t)$.

Continuing this raster match for all TDOA pairs, $t_{a,ij,00}$ will be most likely identified several times as the direct path while all other path combinations $(\mu,\nu) \neq (0,0)$ will at least once be identified as non-direct.

This also holds for multiple uncorrelated sources, as long as their cross-correlation peaks don't overlap. To avoid rejecting a direct path TDOA that accidentally fits to the echo of another source, the decision to reject a TDOA should be made after all direct paths are enhanced and all echo paths are deemphasized in their quality depending on the matching auto-correlation amplitude. As in a practical digital application TDOA differences won't fit exactly due to noise and sampling, we propose to include a narrow smoothing window in order to allow and evaluate approximate raster matching.

Ideally, we get a set of direct path TDOAs $\{t_{a,ij,00}\}$ for each sensor pair (i,j) after the raster match process. It is the task of the next step to assign them to different sources. In practice, the raster match process is not perfect and the set won't contain all direct path TDOAs ("miss"). It might also contain some non-direct path TDOAs ("false alarm"). They will be partly rejected by the following steps.

4.2. Sensor-triple with zero TDOA sum

Observation A2 is trivial for $\tilde{M} = 2$ sensors. It will now be evaluated for a sensor-triple (i,j,k). Increasing the number of involved sensors in the zero TDOA sum condition beyond $\tilde{M} = 3$ will cause ambiguity of the sum expression and increase the computational effort as well as rounding errors in practical applications.

Any triple combination of TDOAs out of sets $\{t_{a_1,ij,\mu_1\nu_1}\}$, $\{t_{a_2,jk,\nu_2\kappa_2}\}$, and $\{t_{a_3,ki,\kappa_3\mu_3}\}$ whose sum disappears, most likely belongs to the same source $a_1 = a_2 = a_3$ and has common paths $\mu_1 = \mu_3$, $\nu_1 = \nu_2$, and $\kappa_2 = \kappa_3$. Comparing all triple TDOA combinations for all $\binom{M}{3}$ sensor-triples (i,j,k), we end up with TDOAs having two or more matching partners, each associated with the same source and paths. Each match increases the quality of all three TDOA partners.

Assuming that every source a is detected by at least 3 sensors – otherwise we could not localize it – we discard all TDOAs with no matching partners. Theoretically, matching TDOA triples of non-direct paths are possible, e.g. due to reflected sources, but they have in general a lower quality due to lower correlation amplitudes and fewer partners and thus might be rejected in the next step.

Again as before, sampled TDOA estimates won't fit exactly, so we propose to include a narrow smoothing window.

4.3. Combining triples to source vectors

In order to obtain all $\binom{M}{2}$ TDOAs for each detected source, TDOA triples are finally combined in the following way: The triple of highest quality initializes the first source vector which contains all TDOAs assigned to one source. We then search for those remaining triples which share one common TDOA with the initial one and assign the remaining two TDOAs to the source vector, see Fig. 2. By continuing the search using the extended source vector, more and more TDOAs will be assigned to this source.

With the subset of remaining, not yet assigned TDOA triples we form the next source vector like before and continue this process until each triple has contributed to at least one source.

On the basis of the number of matching TDOAs, the number of matching triples, and their quality values, we can finally calculate an overall quality for each source vector and use it as a reliability measure for the existence of the corresponding source.

[1] Auto-correlation peaks having no direct path involved may be identified similarly by exploiting $|t_{a,i,\mu\nu}| = |t_{a,i,\mu 0}| + |t_{a,i,\nu 0}|$ and thus are excluded from further considerations.

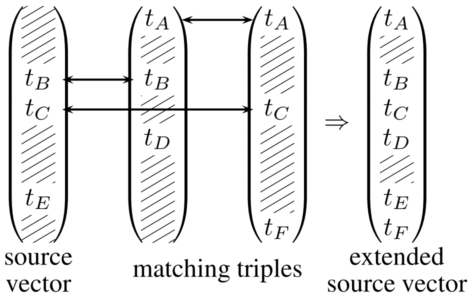

Fig. 2. Illustration of the combination algorithm: hatched area is yet unknown, $t_{A...F}$ stand for any TDOA $t_{a,ij,00}$.

In practical applications the source vectors might be incomplete. For some subsequent algorithms of geometric position estimation, this could have an effect (e.g. for the choice of reference sensor for [4]), for others (e.g. [3]) it doesn't matter. Besides, some TDOAs still may be estimated according to observation A2, if necessary.

5. SIMULATION RESULTS

To illustrate the performance of DATEMM, sensor signals were simulated using the image method [11]. Several different setups were analyzed and show promising results.

In the following example two speech sources and five sensors were randomly placed in a noise-free, small office room having highly reflecting walls ($T_{60} \approx 0.3$s). TDOAs were estimated from a time window of 40 ms, in which both sources were active. In order to obtain high resolution TDOAs, the sampling rate was 96 kHz. Fig. 3 shows the successful disambiguation of the 10 most dominant peaks in the cross-correlation of one sensor pair.

A typical output of DATEMM is presented in Table 1. The algorithm found four possible sources of different quality. Obviously, the first two estimated source vectors correspond well to the true source $a = 2$ and $a = 1$. The more unlikely source vectors are partially due to image sources.

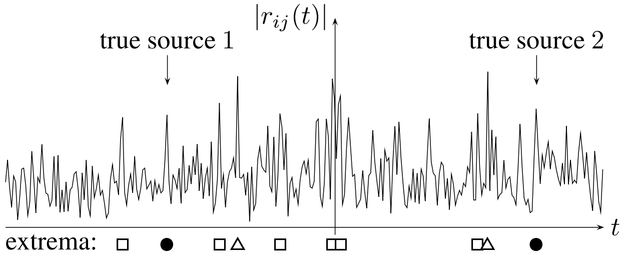

Fig. 3. Example of a TDOA classification into echo path □, non-matching triple △ , and direct path ●

	true TDOA		estimated TDOA			
	$a = 1$	$a = 2$	(source vectors)			
$t_{a,12,00}$	174.7	4.8	5	175	392	392
$t_{a,13,00}$	-192.2	68.3	68	-192	25	n.a.
$t_{a,14,00}$	-364.6	-120.7	-121	-362	n.a.	264
$t_{a,15,00}$	109.8	-324.9	-322	110	327	60
$t_{a,23,00}$	-366.9	63.5	63	n.a.	n.a.	n.a.
$t_{a,24,00}$	-539.3	-125.5	-126	-539	1	-126
$t_{a,25,00}$	-64.9	-329.6	-330	-65	-65	-330
$t_{a,34,00}$	-172.4	-189.1	-190	n.a.	n.a.	n.a.
$t_{a,35,00}$	301.9	-393.1	-391	303	303	n.a.
$t_{a,45,00}$	474.3	-204.2	-204	474	-64	-204
		Normalized quality:	45%	36%	12%	7%

Table 1. Comparison of true TDOAs and DATEMM output

6. REFERENCES

[1] C. Knapp and G. Carter, "The gereralized correlation method for estimation of time delay," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 320–327, 1976.

[2] J. Delosme, M. Morf, and B. Friedlander, "Source location from time differences of arrival," in *ICASSP*, 1980.

[3] R. Schmidt, "A new approach to geometry of range difference location," *IEEE Trans. on Aerospace and Electronic Systems*, vol. 8, pp. 821–835, 1972.

[4] Y. Huang, J. Benesty, G. Elko, and R. Mersereau, "Real-time passive source localization: A practical linear-correction least-squares approach," *IEEE Trans. on Speech and Audio Processing*, vol. 9, pp. 943–956, 2001.

[5] Y. Huang and J. Benesty, "Adaptive multichannel time delay estimation based on blind system identification for acoustic source localization," in *Adaptive Signal Processing*. Springer Verlag, 2003.

[6] J. Chen, Y. Huang, and J. Benesty, "Time delay estimation via multichannel cross-correlation," in *ICASSP*, 2005.

[7] E. Di Claudio and R. Parisi, "Multi-source localization strategies," in *Microphone Arrays: Signal Processing Techniques and Applications*. Springer Verlag, 2001.

[8] D. Sturim, M. Brandstein, and H. Silverman, "Tracking multiple talkers using microphone-array measurements," in *ICASSP*, 1997.

[9] P. Heracleous, S. Nakamura, and K. Shikano, "A microphone array-based 3-D N-best search algorithm for the simultaneous recognition of multiple sound sources in real environments," *IEICE Trans. on Information and Systems*, vol. E85-D, pp. 994–1002, 2002.

[10] Y. Huang, J. Benesty, and J. Chen, "A blind channel identification-based two-stage approach to seperation and dereverberation of speech signals in a reverberant environment," *IEEE Trans. on Speech and Audio Processing*, vol. 13, pp. 882–895, 2005.

[11] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, 1979.